Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Informationsextraktion aus semi-strukturierten Befundberichten

Martin Toepfer

Arbeitsgruppe zum Projekt „Data Warehouse“:
Prof. Dr. Frank Puppe
Philip-Daniel Beck, Georg Dietrich, Maximilian Ertl,
Georg Fette, Dr. Mathias Kaspar, Jonathan Krebs

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Universität Würzburg

TMF-Workshop "Textmining für die medizinische Forschung - wie weit sind wir? (Berlin, 28.01.2015)

Projektübersicht

- klinisches Data Warehouse für die medizinische Forschung
  - Patientenrekrutierung für Studien
  - Patientendatenbereitstellung für Studien
  - Statistik und Data Mining …
- viele **wichtige Informationen liegen nicht strukturiert vor**
- vorhandene Terminologien genügen nicht den Anforderungen
  - z.B. Spezialterminologien für Studien in Echokardiografie notwendig

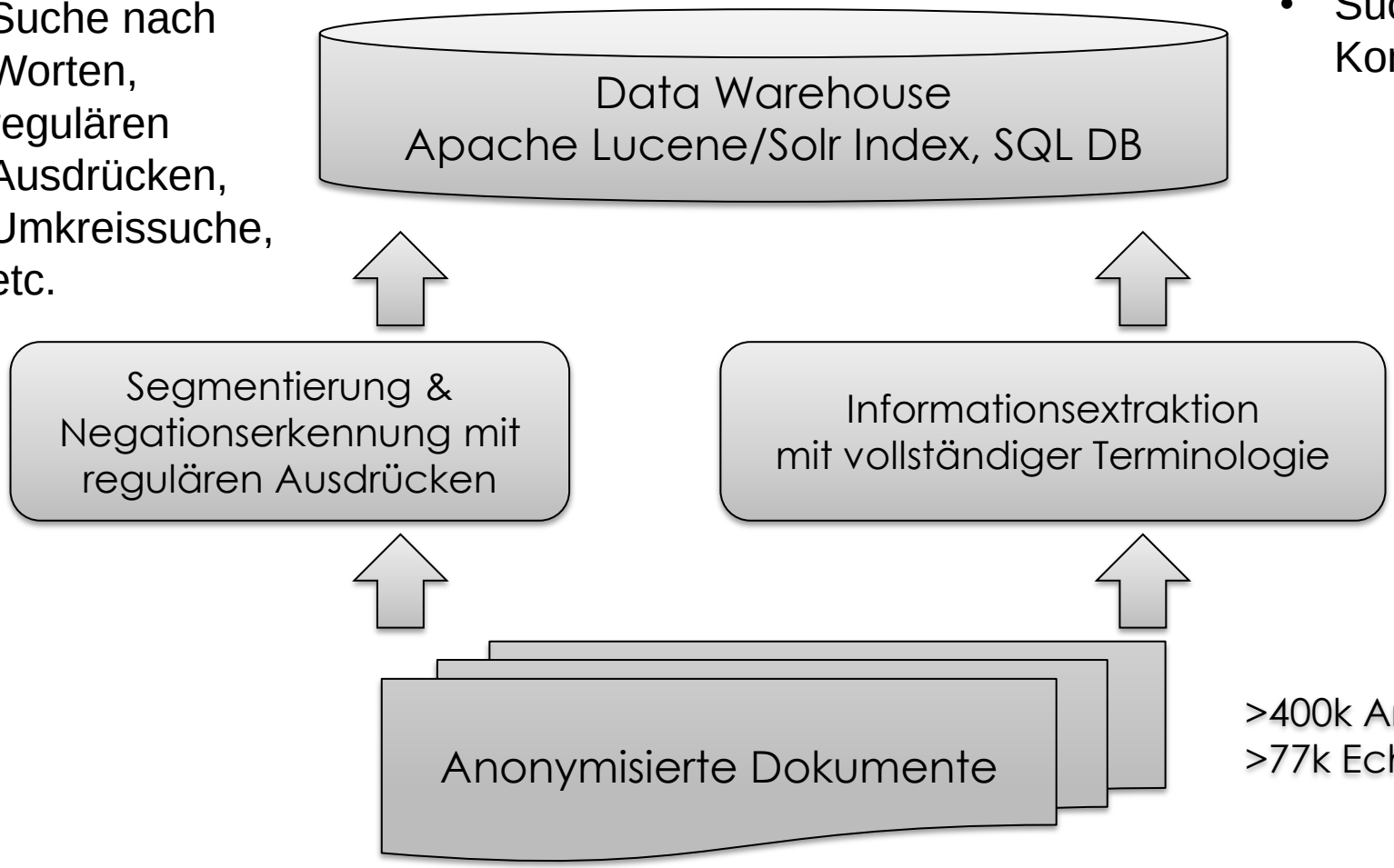
Ansätze zur Erschließung textueller Dokumente

„ad hoc“ (~IR):

- Suche nach Worten, regulären Ausdrücken, Umkreissuche, etc.

„offline“ (~IE):

- Suche nach Konzepten



>400k Arztbriefe
>77k Echos

Merkmale der Ansätze

	„ad hoc“ (mit retrieval Techniken)	**„offline“** (information extraction)
vollständige Terminologie	nein	ja, daher Data Mining über Konzepten möglich
Auflösung von Mehrdeutigkeit	Begrenzt möglich, z.B. durch Bereichsanfragen	ja
Entwicklungsaufwand	**	****
Beispielanfrage über das Data Warehouse	„Mitralinsuffizienz MI Mitralklappen-insuffizienz …“	„Mitralklappeninsuffizienz = vorhanden“

Vorverarbeitung

Eingabe: (*.doc/ *.docx/ *.xml)	HTML-Konverter, zweistufige Anonymisierung	Abschnitts-erkennung mit regulären Ausdrücken, Regeln	Selektion bestimmter Abschnitte, z.B., „EKG-Abschnitte“

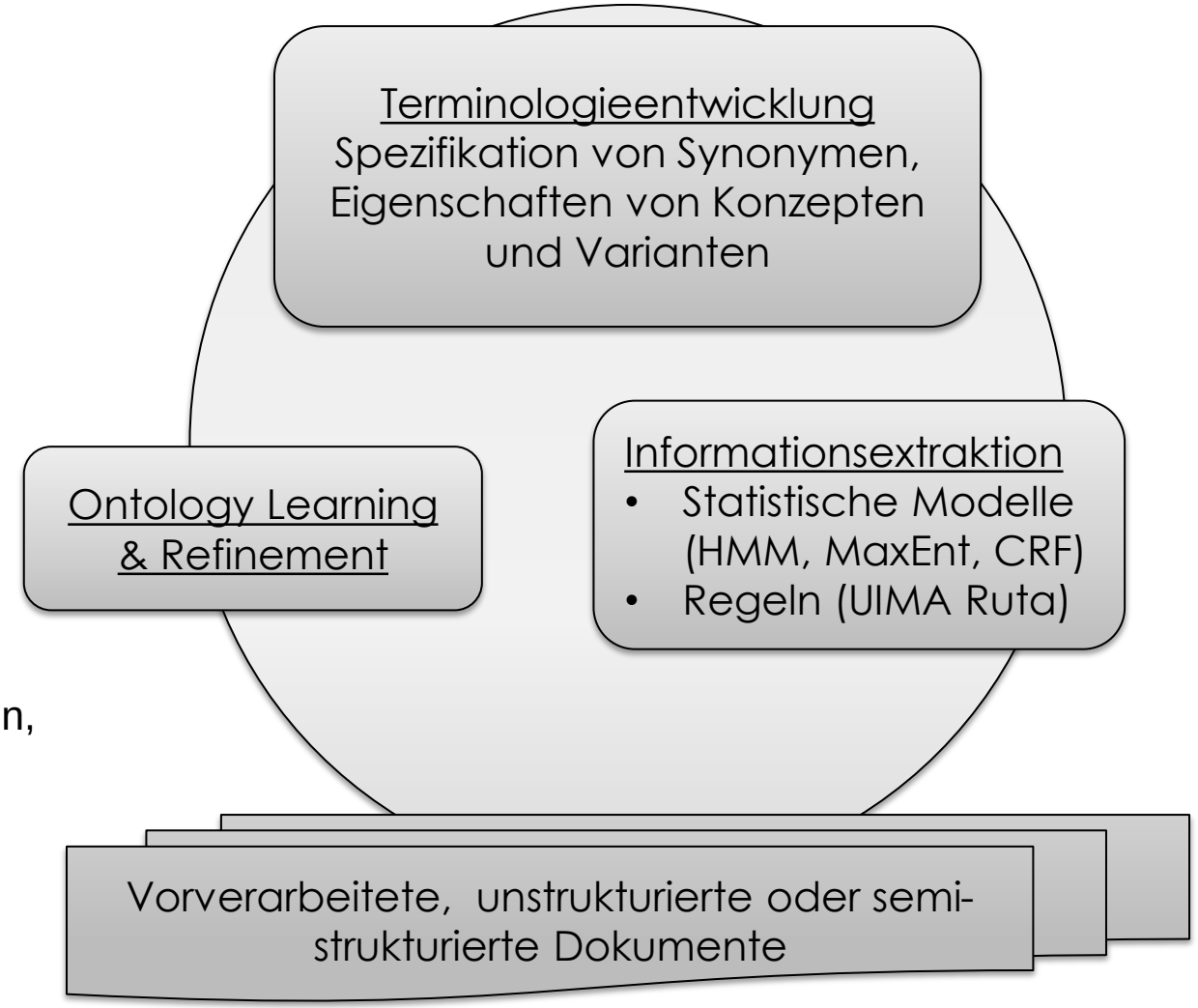
Terminologie-/ IE-Erstellung: allgemeine Übersicht

Domänenexperte

Wissenseingabe

Technikexperte

Auswahl von Komponenten, Anpassung von Regeln



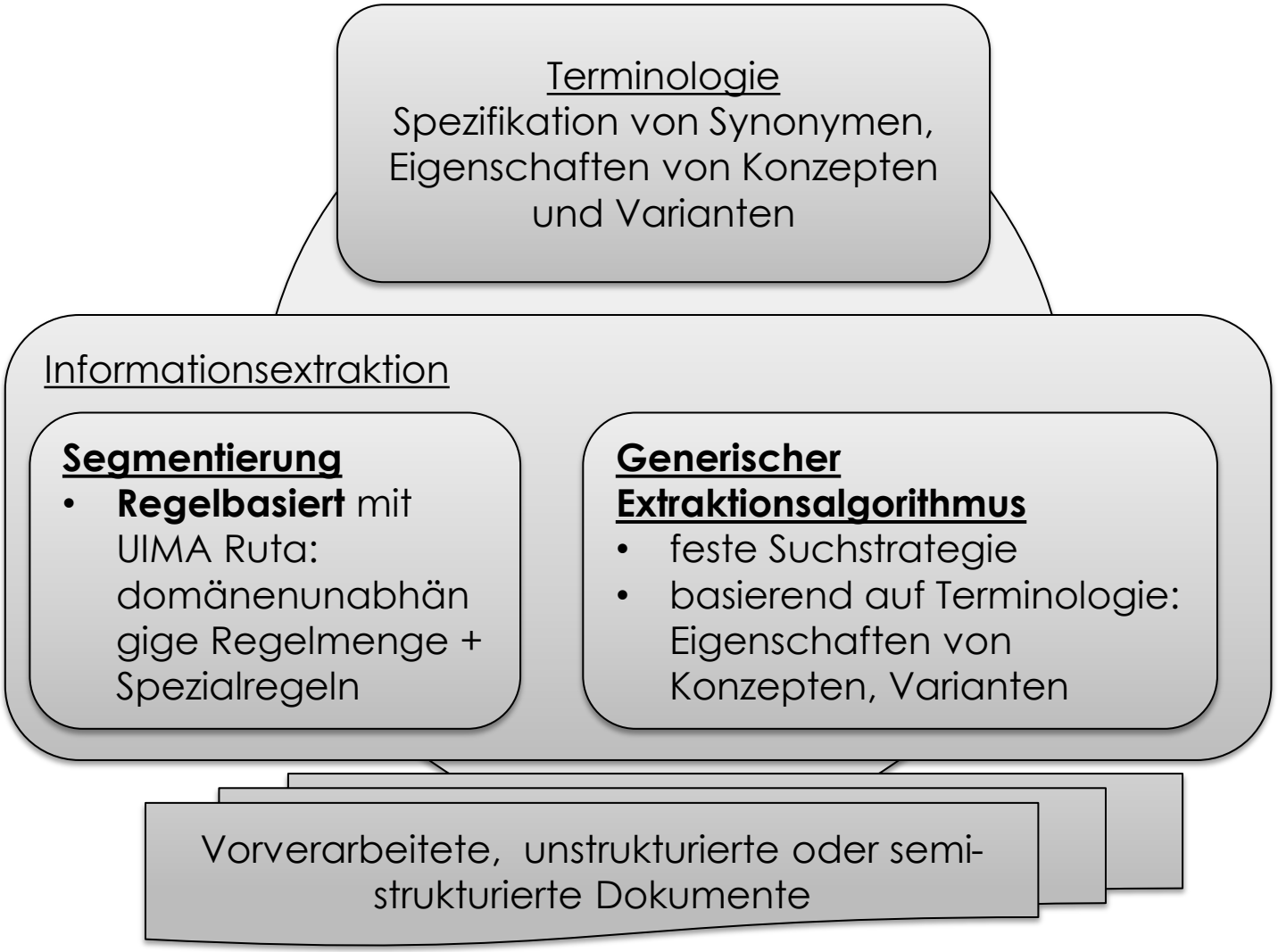
Terminologie-/ IE-Erstellung: semi-strukturierte Dokumente

Domänenexperte

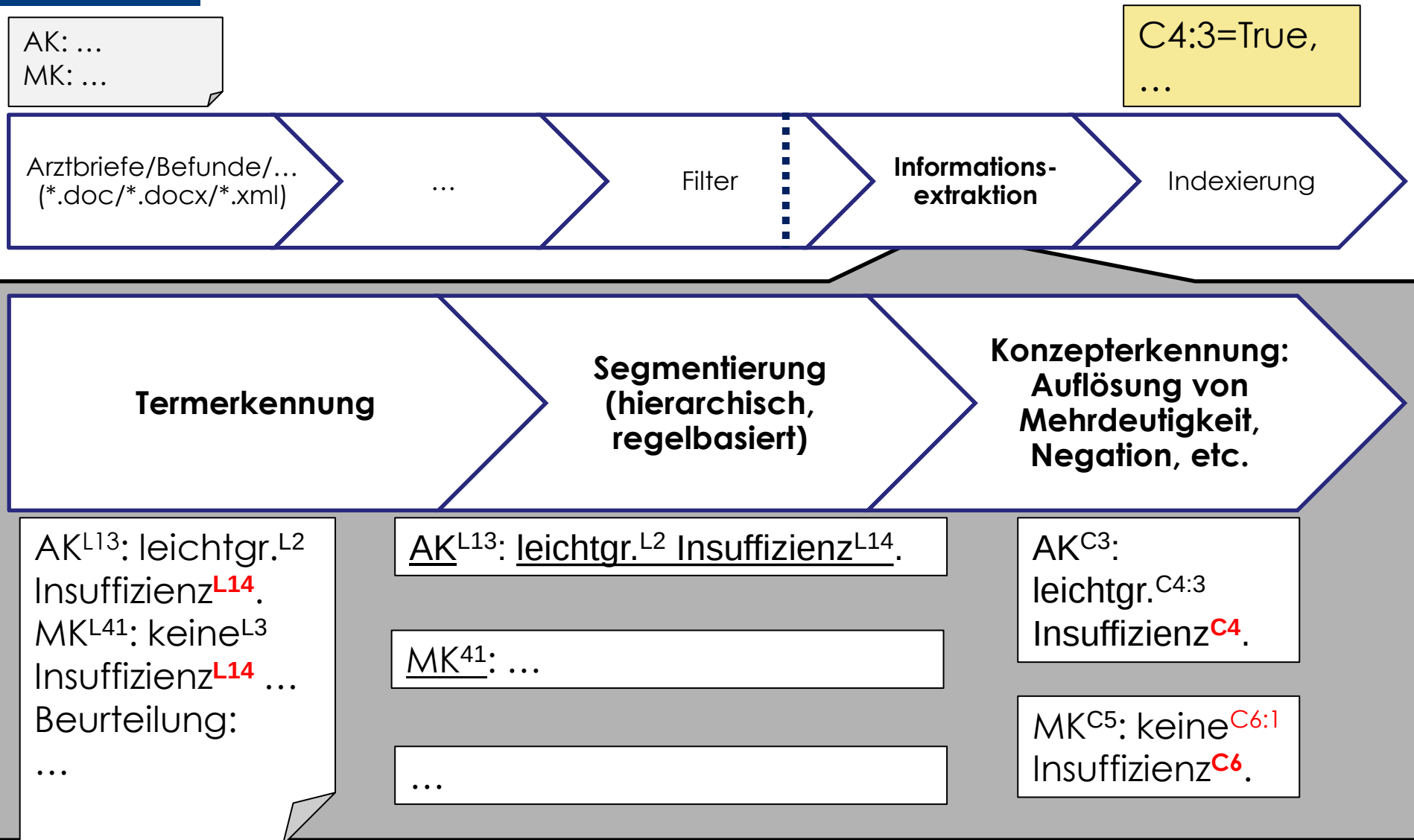
Wissenseingabe

Technikexperte

Anpassung von voreingestellen Standardskripten, -regeln



Dokumentenverarbeitung: Informationsextraktion

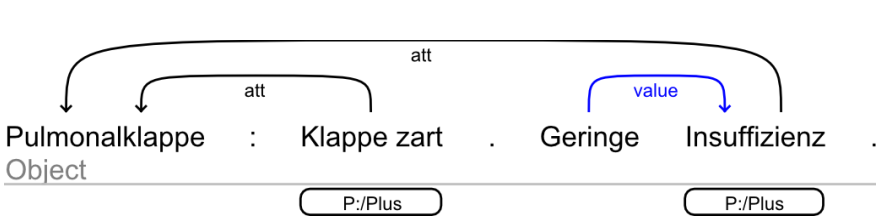


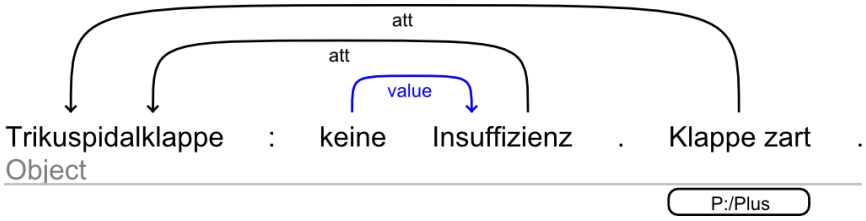
Terminologieeinträge und -eigenschaften

	Beschreibung	**Beispiel**
Objekt	Bezug für Attribute	Aortenklappe
(Objekt-)**Attribut** (+)	Eigenständiges oder an Objekt gebundenes Attribut	Aortenklappeninsuffizienz (obj. attribute), LVEF (attribute)
Wert	Numerischer oder boole'scher Wert eines Attributes	ja, nein, leichtgradig, Frequenz_in_Hz
Template	Menge von Werten semantisch ähnlicher Attribute	*Schweregrade: leicht,* … Stenose (T:Schweregrade), MI (T:Schweregrade)
Variant (std/regex)	Einzelner lexikalischer Ausdruck; einfach oder regulärer Ausdruck; optional: *Objekt-Attribut* oder *Attribut-Wert* Kompositum	„Insuffizienz“ „(\\d+)\\s*mmHg“
Dictionary	Menge von Varianten	Verdacht auf: „Verdacht auf“, „v.A.“, …
Property	Weitere flexible Eigenschaften	MI (ICD-10=„I34.0“)

Beispiel: (vereinfacht) Klassifikation

Konzeptauflösung & Zuweisung von Bezügen:



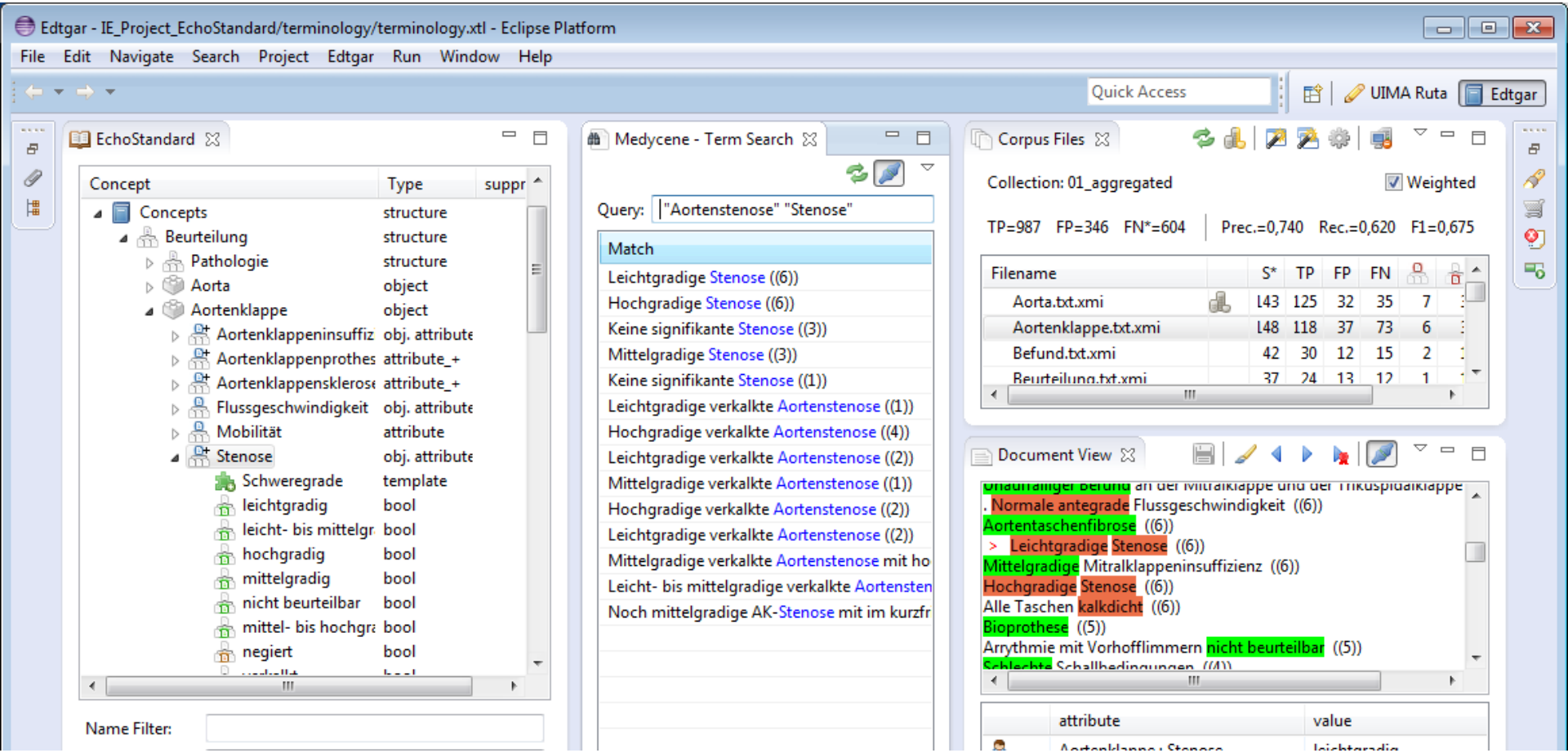


Inferenz:

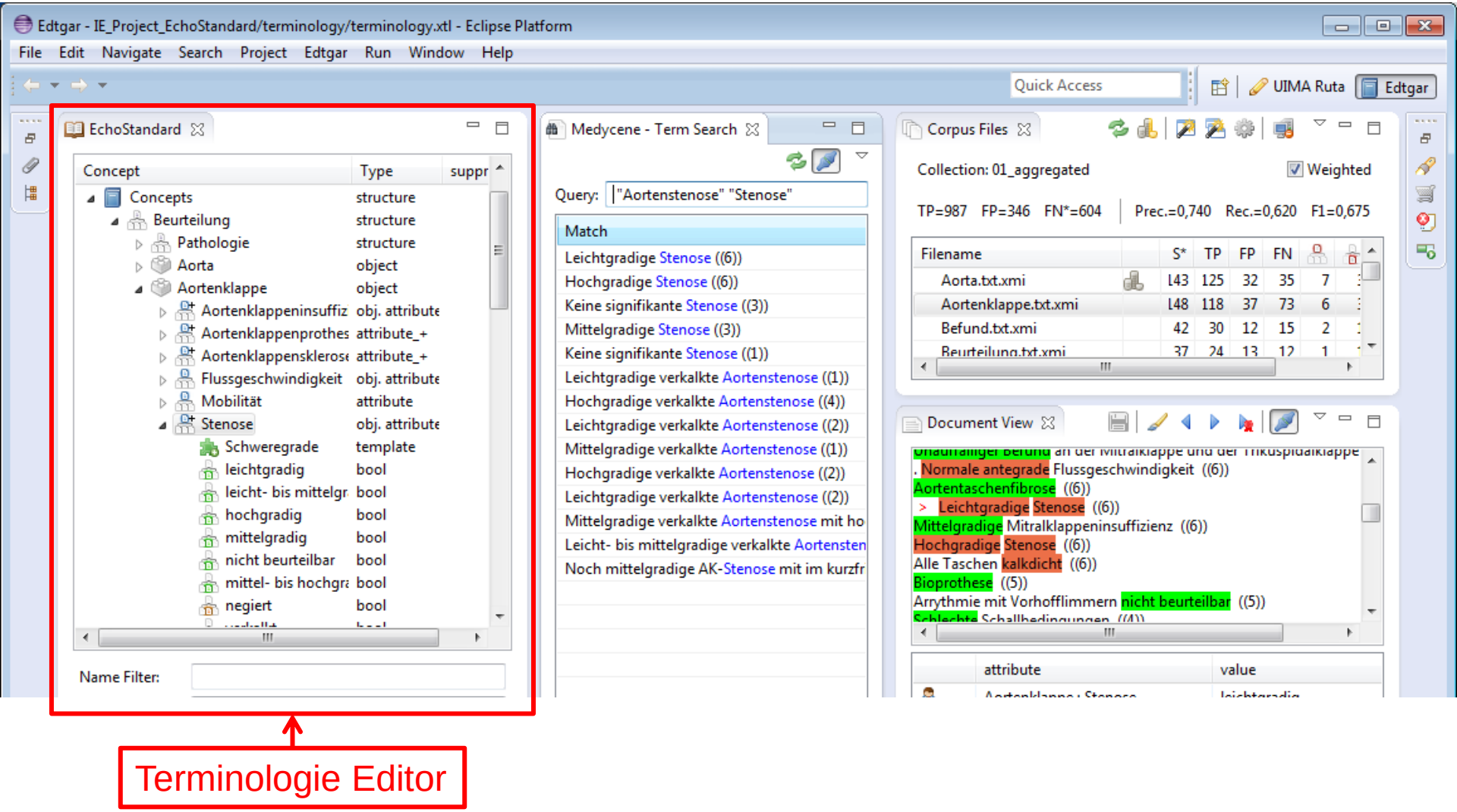
… in V1-V4. …keine konzentrische… …Linksherzhypertrophie…

V1,V2,V3,V4 keine ~~konzentrische~~ Hypertrophie des linken Myokards

Arbeitsumgebung & Tooling

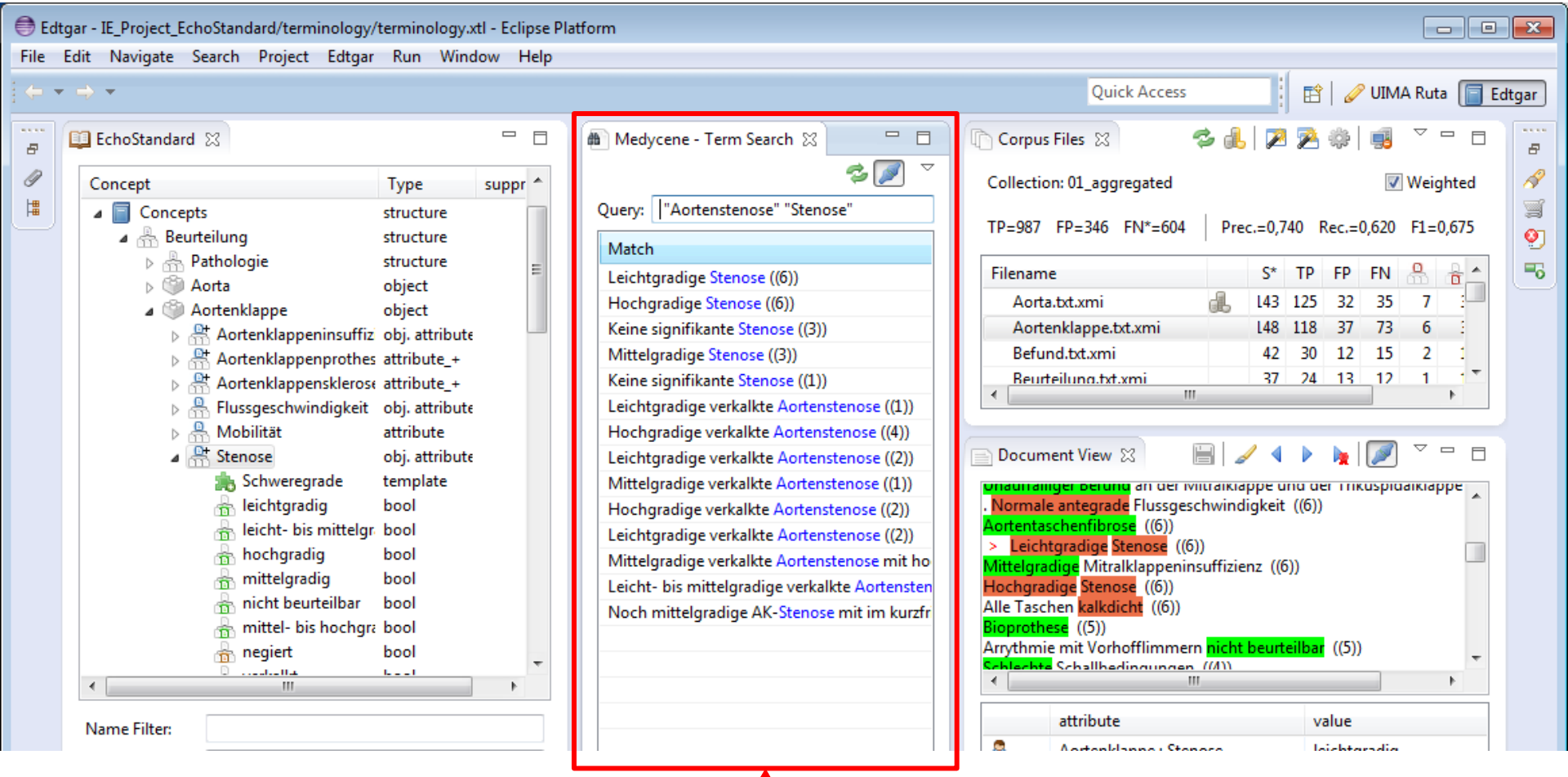


Arbeitsumgebung & Tooling



Terminologie Editor

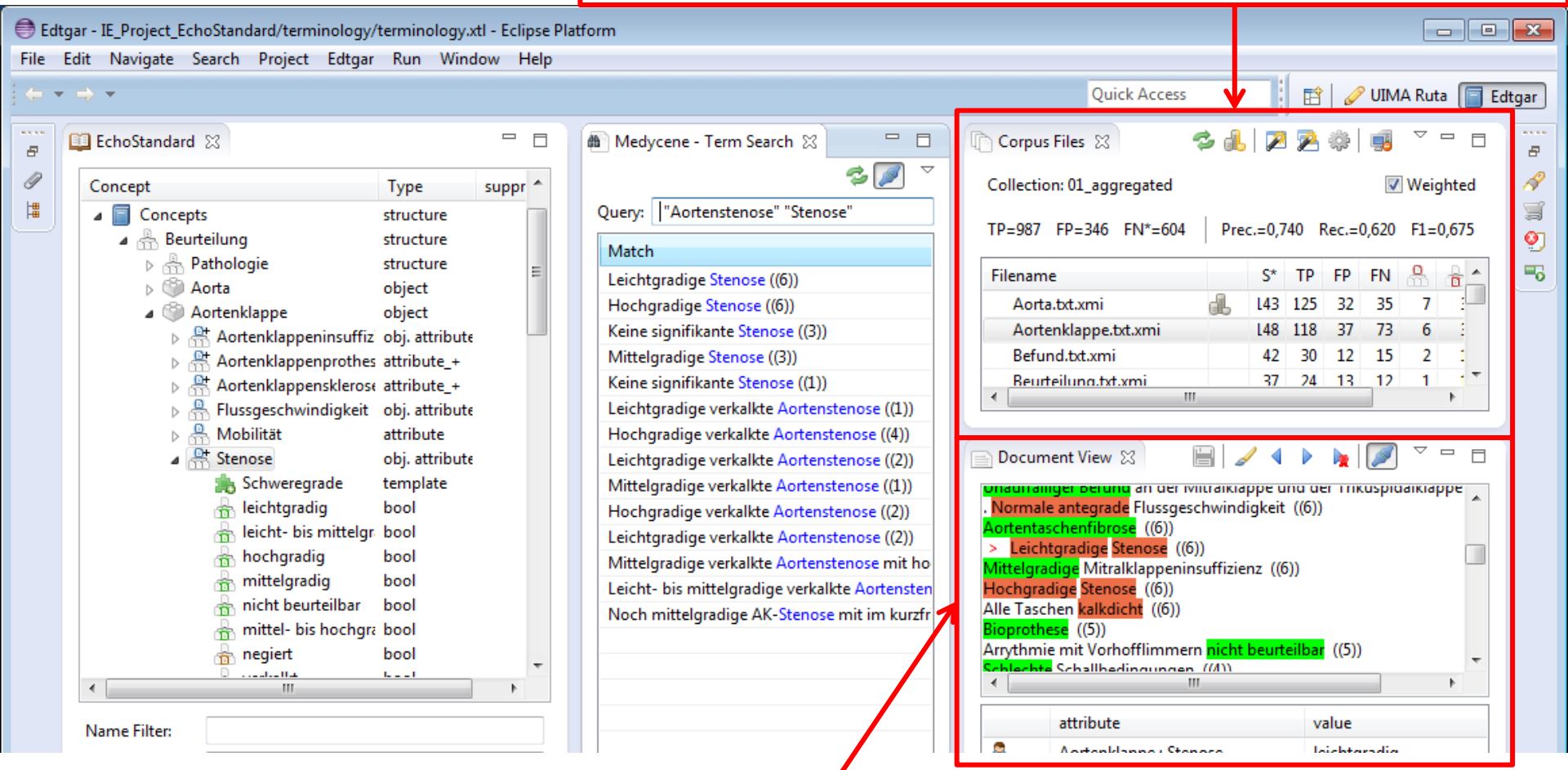
Arbeitsumgebung & Tooling



Semantische- und Freitextsuche über Apache Lucene

Arbeitsumgebung & Tooling

Übersicht zu Extraktion auf Trainings- / Testdatensatz



Dokumentenansicht und
semi-automatische Goldstandarderstellung

Semi-strukturierte Dokumente

- Anordnung und Formatierung kennzeichnen Art und Zugehörigkeit der Informationen
- Domänen:
  - Echokardiografieberichte,
  - Medikationsabschnitte in Arztbriefen,
  - …

Echokardiografie vom 28.01.2015:

Patient: M. Toepfer

Aortenklappe: normal. Keine Insuffizienz.
Mitralklappe: leichtgradiger Rückfluss.
Trikuspidalklappe: …

Gewicht 99kg, Größe: 199cm
Datum: 24.12.2016

- AK: normal. Leichtgradige Insuffizienz.
- MK: normal.
- TK: …

Beurteilung: …

Sehr geehrter Herr Toepfer,

…

Medikation:

Vitamin D	½-0-½
Kaffee	1-1-1

…

Beurteilung: …

Fallstudie – GMDS2014

- Sample für Evaluation:
  200 anonymisierte Echokardiografiebefunde
- Aus dem Gesamtdatenbestand 2012-2013:
  ~21 700 Dokumente
- ca. 50 Attribut-Wert-Paare pro Befundbericht
- (micro-avg.) Precision, Recall, **F1**
  99,4%, 93,2%, **96,2%**
- Fehlerquellen:
  - fehlende Konzepte, Varianten
- Recall leicht verbesserbar durch mehr Dokumente

Zusammenfassung

- Motivation:
  - schnelle Entwicklung von Terminologien und Informationsextraktionsmodulen für klinische Subdomänen
  - einfach einzusetzendes System
- Bisherige Ergebnisse
  - Transthorakale Echokardiografie: sehr hohe Precision, hoher Recall
- In Arbeit: Medikation, EKG, KU, …